АВТОМАТИЧНЕ ВИЗНАЧЕННЯ СЕМАНТИЧНО БЛИЗЬКИХ КОЛОКАЦІЙ АНГЛІЙСЬКОЇ МОВИ

Кисільова В.Ю., Петрасова С.В.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60
e-mail: *b1acksmi1e@mail.ru, svetapetrasova@gmail.com*

Колокація – словосполучення, що має ознаки синтаксично та семантично цілісної одиниці, у якій вибір одного із компонентів здійснюється за змістом, а вибір другого залежить від вибору першого [1].

Колокації, на відміну від окремих слів, яким властива багатозначність та синонімічність, містять у собі більш конкретну семантичну інформацію, тому їх автоматичне визначення є актуальним завданням у галузі автоматичної обробки природної мови.

Ідентифікація колокацій у тексті базується на виявленні синтагматичних відношень у природній мові. У цьому напрямку існують два основних підходи до вивчення синтагматичних відношень. Широко розповсюдженим є *синтаксичний підхід*, у якому сполучуваність колокатів визначається їх сумісністю у словосполученні та/або конкретною синтаксичною моделлю. Цей підхід базується на використанні синтаксичних аналізаторів (парсерів), які допомагають встановити зв'язки між словами у реченні. Другий підхід, що полягає у виявленні статистичних закономірностей при побудові тексту, називається *статистичним.* В основі статистичного апарату виявлення колокацій лежать так звані «міри асоціації», які є показником сили синтагматичного зв'язку між елементами колокацій [2].

На основі цих підходів застосовують наступні методи визначення синонімів та синонімічних колокацій:

– вимірювання семантичної подібності між парами слів через аналіз результатів запитів у пошуковій web-системі [3];

– математичний аналіз слів та їх тлумачень у одномовному словнику;

– визначення подібності слів за допомогою Dice measure [4];

– вимірювання подібності слів через математичне визначення подібності їх перекладів;

– виявлення синонімічних колокацій на основі порівняння їх перекладів [5];

– знаходження перефразувань за подібними фрагментами речень [6];

– аналіз корпусу паралельних перекладів англійських текстів за допомогою математичного визначення подібності контексту [7];

– формалізація поняття семантичної еквівалентності колокацій засобами семантичної та граматичної характеристик колокатів [8].

Більшість з вищенаведених методів характеризуються тим, що для кожної мовної одиниці будується вектор, значення якого характеризують міру семан-

тичної близькості. Однак аналіз існуючих методів показав, що для більш точного визначення семантично близьких мовних одиниць необхідне комплексне використання підходів і методів.

В роботі для виявлення семантично близьких колокацій англійської мови було використано логіко-лінгвістичну модель автоматичної ідентифікації семантичної подібності колокацій, що базується на застосуванні методів штучного інтелекту та множини граматичних і семантичних характеристик для формалізації мовних одиниць [8].

На основі досліджуваної моделі були розроблені наступні правила формування колокацій англійської мови:

$$a_{y1}^{\text{Noun Obj}} c_{y1}^{\text{Att}} a_{x1}^{\text{Noun Sub}} c_{x1}^{\text{Ag}} \approx a_{y2}^{\text{Noun Obj}} c_{y2}^{\text{Att}} a_{x2}^{\text{Noun Sub}} c_{x2}^{\text{Ag}} \approx \approx a_{x3}^{\text{Noun SubOf}} c_{x3}^{\text{Ag}} a_{y3}^{\text{Noun Obj}} c_{y3}^{\text{Att}}, \quad (1)$$

$$a_{x1}^{\text{Noun Sub Pl}} c_{x1}^{\text{Ag}} a_{y1}^{\text{Verb}} \approx a_{x2}^{\text{Noun Sub Pl}} c_{x2}^{\text{Ag}} a_{y2}^{\text{Verb}} \mid a_{x1}^{\text{Noun Sub Sing}} c_{x1}^{\text{Ag}} a_{y1}^{\text{Verb+s/es}} \approx \approx a_{x2}^{\text{Noun Sub Sing}} c_{x2}^{\text{Ag}} a_{y2}^{\text{Verb+s/es}}, \quad (2)$$

$$a_{x1}^{\text{Verb}} a_{y1}^{\text{Noun Obj}} c_{y1}^{\text{Pac}} \approx a_{x2}^{\text{Verb}} a_{y2}^{\text{Noun Obj}} c_{y2}^{\text{Pac}}, \quad (3)$$

$$a_{y1}^{\text{Adj Att}} a_{x1}^{\text{Noun Sub}} c_{x1}^{\text{Ag}} \approx a_{y2}^{\text{Adj Att}} a_{x2}^{\text{Noun Sub}} c_{x2}^{\text{Ag}} \approx a_{x3}^{\text{Noun Sub}} c_{x3}^{\text{Ag}} a_{y3}^{\text{Adj Pr}}, \quad (4)$$

де x_i – головне слово колокації; y_i – залежне слово колокації; граматичними характеристиками є: $a^{\text{Noun Sub Pl}}$ – іменник множини, суб'єкт; $a^{\text{Noun Sub Sing}}$ – іменник однини, суб'єкт; $a^{\text{Noun Obj}}$ – іменник, об'єкт; $a^{\text{Adj Att}}$ – прикметник, атрибутивний; $a^{\text{Adj Pr}}$ – прикметник, предикативний; a^{Verb} – дієслово множини; $a^{\text{Verb+s/es}}$ – дієслово однини; та семантичні характеристики: c_{xi}^{Ag} – агенс; c_{yi}^{Att} – атрибут; c_{yi}^{Pac} – пацієнс.

В результаті визначення множини семантико-граматичних характеристик колокатів було розроблено наступний алгоритм автоматичного формування семантично близьких колокацій.

На першому етапі, після того як користувач вводить колокацію, програма виконує пошук її колокатів у базі даних. У разі знаходження введених користувачем колокатів за внутрішнім зв'язком у базі даних визначаються синоніми до кожного (головного та залежного) слова вхідної колокації.

Наступним кроком є перевірка відповідності граматичних та семантичних характеристик між колокатами введеної користувачем колокації та знайденими синонімами. За умови успішної перевірки програма будує синонімічне словосполучення згідно розроблених правил (формул).

Наприклад, семантична еквівалентність колокацій: *a baud rate* $\approx$ $\approx$ *a transfer speed* $\approx$ *a speed of transfer* визначатиметься згідно формули 1, а таких колокацій як *to store data* $\approx$ *to keep information* – згідно формули 3.

Таким чином, у результаті розроблена програма відображає семантично близькі колокації англійської мови.

Список літератури

1. Manning Christopher D. Foundations of Statistical Natural Language Processing / Christopher D. Manning, Hinrich Schütze. – MIT Press, Cambridge, 1999. – 680 pp.

2. Захаров В.П. Анализ эфффективности статистических методов выявления коллокаций в текстах на русском языке / В.П. Захаров, М.В. Хохлова // «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011. – Воронеж, 2011. – с. 134–138.

3. Nakov S. Automatic Acquisition of Synonyms Using the Web as a Corpus / S. Nakov // Proceedings of the 3rd Annual South-East European Doctoral Student Conference (DSC 2008). – Vol. 2. – P. 216–229.

4. Hua Wu Optimizing Synonym Extraction Using Monolingual and Bilingual Resources / Hua Wu, Ming Zhou // Proceedings of the second international workshop on Paraphrasing (PARAPHRASE '03). – Stroudsburg, PA, USA, 2003 – Vol. 16. – P. 72–79.

5. Hua Wu Synonymous Collocation Extraction Using Translation Information / Hua Wu, Ming Zhou // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03). – Stroudsburg, PA, USA, 2003 – Vol. 1. – P. 120–127.

6. Marius P. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web / P. Marius, D. P´eter // Proceedings of the Second International Joint Conference: Natural Language Processing (IJCNLP 2005). – Korea, 2005. – P. 119–130.

7. Barzilay R. Extracting Paraphrases from a Parallel Corpus / R. Barzilay, Kathleen R. McKeown // Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01). – Stroudsburg, PA, USA, 2001. – P. 50–57.

8. Khairova N. The logical and linguistic model for automatic extraction of collocation similarity / N. Khairova, S. Petrasova, Ajit Pratap Singh Gautam // Econtechmod : an international quarterly journal on economics in technology, new technologies and modelling processes. – Lublin; Rzeszow, 2015. – № 3 (4). – P. 43–48.